EXHIBIT A177

Annu. Rev. Public Health. 1989. 10:423–40

PERSPECTIVES ON STATISTICAL SIGNIFICANCE TESTING

Robert F. Woolson

University of Iowa, Department of Preventive Medicine, Iowa City, Iowa 52242

Joel C. Kleinman

National Ccntcr for Health Statistics, Hyattsville, Maryland 20782

Introduction

In this paper we summarize some of the issues surrounding the "significance testing" controversy. We present many of the points that have been raised as criticisms against the use of statistical hypotheses testing. We hope that such discussion will lead to a greater understanding of the role that statisitical inference, of several varieties, plays in the analysis of biomedical data. Our position is that both confidence interval estimation and significance testing have major roles to play in the analysis of medical and public health data.

We begin by describing this controversy as it has emerged recently and discuss its evolution in several related fields, particularly psychology and sociology. The discussion in these fields closely parallels the discussion taking place in the biomedical field. Our review of the historical issues and discussions in related fields is not intended to be comprehensive or all inclusive, but we believe it captures the main points. After this discussion we present a summary of some of the arguments for and against significance testing. We conclude with an example that emphasizes the importance of the use of both confidence interval estimation and significance testing for the analysis of epidemiologic data.

Background

Over the last 10 or 11 years, and in particular the past 3 years, the field of public health has witnessed a lively debate on the use of significance testing in public health and medical research. Rothman (38) criticized the use of

0163-7525/89/0510-0423$02.00

significance testing and reliance on p-values for the analysis of medical data. Contained in this article is a strong recommendation that confidence intervals not only be used for the analysis of biomedical data, but that they be used entirely in place of statistical tests of significance. Rothman and his colleagues (34–36) continue to foster the view that statistical significance testing is of little value for the analysis of biomedical data and apparently would like to limit the use of such tests. This view has also been expressed by others, including Gardner & Altman (18), and has also been presented in rather strong editorial positions, e.g. the *American Journal of Public Health* editorial comment by Rothman & Yankauer (39). Although all these papers do not necessarily explicitly state that significance testing be banished entirely, there is no question that the view expressed in these papers is that significance tests be used only in the most limited settings. This position could have a strong impact on public health research, since many epidemiologists and statisticians use significance tests routinely for analysis of their data, especially complex data sets.

Others (41, 48, 49), have also endorsed the position that significance tests be reserved for use in the most limited settings and that confidence interval estimation and descriptive estimation be the primary statistical tools for data analysis. Fleiss (9–11) has been the most vigorous proponent for the use of statistical significance tests for the analysis of public health and medical data, and in our view points out a number of situations in which the use of such tests is beneficial, perhaps absolutely essential. He describes selected circumstances in which significance tests have specific roles for the analysis of epidemiologic data, and also explicates the difficulties that could be encountered by the abolition of the concrete decision rules defined by statistical significance tests. He also carefully emphasizes situations in which statistical significance tests would be inappropriate, and acknowledges that there has been considerable misuse of both significance testing and statistical procedures in general in health research.

Thompson (44, 45) also delineates guidelines for the presentation and analysis of data, and outlines circumstances in which confidence interval estimation and significance testing should be used. He emphasizes the reporting of precise p-values in place of degrading the results of a significance test into a simple "accept/reject" categorization. He emphasizes, however, the importance of using confidence interval estimation as a primary tool for statistical analysis, but notes that p-values can and should be reported in many data analyses.

As part of the ongoing debate on the use of significance testing and confidence interval estimation, a number of papers, including Mills et al (25) and Moses et al (28, 29), have been subjected to critical review by the proponents of one position or the other. Points in question have included whether tests should be used at all or in preference to confidence interval estimation and deal as well with the issue of how one determines whether an association is meaningful without the use of significance testing.

Freiman et al (17) and Bandt & Boen (2) have discussed several other matters regarding the use of significance testing for the analysis of biomedical data. These authors emphasize the role that sample size plays in the Type II error of a statistical significance test. They also stress the importance of distinguishing between a statistically significant result and a substantively significant result.

Diamond & Forrester (7) and Browner & Newman (4) have also contributed to the discussion related to the interpretation of p-values, and mention the important role that Type II errors play in data analysis. Both invoke Bayes' Theorem to introduce a posterior probability associated with the result of a significance test. Relationships are also described between significance testing and the problem of diagnostic sensitivity/ specificity associated with a screening procedure.

The debate involving the use of p-values in medical research has resulted in a variety of guidelines proposed for the reporting of medical data. Gardner & Altman (18) clearly emphasize the importance of confidence intervals and recommend their use in place of p-values. Vaisrub (46) and Bailar & Mosteller (1) also develop guidelines for the reporting of statistical information in medical journals. Bailar & Mosteller (1) emphasize the importance of confidence interval estimation and also highlight the importance of carefully reporting results associated with statistical significance testing. Their guidelines include a strong recommendation to report the exact p-value rather than a simple "significant/nonsignificant" categorization, which has been criticized by many on both sides of the controversy. Many, including us, would agree with this recommendation.

Historical Perspective

Berkson (3) was one of the first to question the uncritical application of statistical tests of significance, particularly with regard to their use as a strategy in the process of accumulating scientific knowledge. He describes a number of settings in which "testing logic" seems to be at odds with scientific logic and how new information is accumulated into scientific thought. In statistical hypothesis testing, one often sets as the null hypothesis the hypothesis that he/she hopes to "disprove." Under the assumption that this null hypothesis is true, one then observes a statistic, and if this statistic has a very small probability of occurrence under the null hypothesis, the null hypothesis is rejected. Berkson (3) criticizes the use of hypothesis testing from two vantage points. First, as noted above, he questions whether testing is in basic disaccord with scientific thought. Second, Berkson feels that the usual application of tests does not stress the importance of specifying alternative hypotheses.

With regard to the first point, Berkson states that the application of significance testing could be described as follows: "If A is true, B will sometimes happen; therefore, if B has been found to happen, A can be considered disproved." Although this is undoubtedly an overstatement of the use of null hypothesis testing procedures, it does raise an interesting logical point.

This point of logic is somewhat troubling to Berkson, but perhaps the more troubling point is the second, related one, namely the failure to specify meaningful alternative hypotheses. Rather than emphasizing rejection of hypotheses for the occurrence of infrequent events, Berkson argues that one ought carefully to specify alternative hypotheses for which the event would have a frequent likelihood of occurrence.

Berkson cites an example described by Fisher (12) that involves a linear regression of a dependent variable (the number of eye facets of *Drosophila melanogaster*) on an independent variable (temperature). In this analysis Berkson reproduces the data described by Fisher. The plot he presents is based on rather large sample sizes. It gives very strong graphical evidence of a linear relationship between the mean facet number and temperature. In spite of this marked visual impression, the application of a statistical significance test leads Fisher to conclude that there is sufficient evidence to reject the hypothesis that the data are linear. Berkson (3) is astonished by this point, noting particularly that the data are as linear as one could hope to find with biologic data. Berkson argues convincingly that the data are consistent with linearity, but are inconsistent with the regression assumption that the temperatures (i.e. the independent variable) are measured without error. This, of course, is one of the key assumptions in linear regression analysis. With such errors of measurement in the independent variable the apparent departure from linearity could be explained by the failure to meet this assumption. As Berkson points out, at least such a hypothesis offers an alternative explanation for the data rather than merely rejecting a hypothesis of linearity, when in fact the data appear to be strongly linear by usual graphical representation. Thus, Berkson is critical of automatic application of null hypothesis testing. He encourages the search for hypotheses for which the data are likely, not the mere rejection of those hypotheses for which the data are not. In short, he encourages proper application of the procedures, not their abolition.

In addition to the above, Berkson also encourages the study of so-called middle p-values (i.e. between 0.3 and 0.7). With adequate sample sizes he feels that such p-values offer evidence in favor of null hypotheses; however, he also stresses the study of such p-values for specific alternative hypotheses. The thesis of his argument is that scientific data should be viewed as being positive for some particular hypotheses, null or otherwise. Data should be interpreted in this light, rather than in the negative light in which traditional statistical hypothesis testing would characterize it. The study of middle p-values, or p-values in general, is similar to the recent suggestion of plotting p-value functions [e.g. Poole (35)].

Berkson (3) is critical of an overemphasis on significance testing in the teaching of statisitical methods. He feels this may be in part due to how mathematical statistics and statistical inference evolved historically. He noted that Fisher's seminal works (12–16) had a major impact on the development of statistical thought in this century. In addition, the work of Neyman & Pearson (30, 31) and Pearson (32, 33) have contributed greatly to the development of statistical hypothesis testing. All of this work was swept up into Wald's 1950 text (47) that cast many statistical problems in the context of statistical decision functions. No doubt these decision procedures have had important application in the area of acceptance sampling for defectives when one is inspecting quality of production material. Berkson argues that such procedures may not be necessarily directly applicable to general scientific research work. The applications studied by Wald were those in which Type I and Type II errors could be clearly defined and the cost incurred with each of these errors could be adequately quantified. In many scientific endeavors, particularly when one is trying to understand association between many variables or when the degree of control varies greatly in the studies performed, one may be unable to quantify the losses or costs incurred with Type I and Type II errors. In addition, scientists rarely, as noted by Berkson (3), dichotomize the results of their work into accepting or rejecting hypotheses but seek other data that support or refute the findings of the specific investigation in question. Theories are generated that provide likely explanation for experimental outcomes. Berkson does not recommend that statistical hypothesis testing be purged from statistical inference; he encourages a more critical use of the method. He recommends the study of middle p-values, and the careful examination of hypotheses that are supportive of the conclusions seen in the data.

To emphasize the role of sample size in interpreting the results of significance tests, and to emphasize the importance of concluding in favor of the null hypothesis rather than merely stating insufficient information to reject the null hypothesis, Berkson offers an example. In Table 1 are the hypothetical results of a physician's judgment based on some serological test, designed to ascertain the sex of a fetus in utero. For each of two situations the probability of obtaining a result as good as the one obtained is 0.38 when the null hypothesis is true. The null hypothesis here is that the serological test is no better than guessing. Do we conclude the same thing with these two data sets? Clearly, we do not. Berkson (3) notes: "Experience 2, being based on large numbers, is convincing underlying positive evidence of the truth of the null hypothesis." With Experience 1, Berkson states, "We cannot say anything from this experience; it certainly does not present any convincing evidence that the physician can discriminate between the sexes. But I would not want to say either that he cannot discriminate. The experiment is too small for any conclusions." Hence, for equal *p*-values, but considerably different sample sizes, one concludes considerably more with Experience 2 then with Experience 1.

It should be noted that this example emphasizes the role that confidence intervals and *p*-values can jointly serve in data analysis. The 95% confidence interval for the proportion correct is 0.3 to 0.9. for Experience 1, but is 0.47 to 0.54 with Experience 2. At best with *p*-values and/or confidence intervals, Experience 1 is inconclusive. Berkson's point is that when numbers are small, a middle *p*-value (0.3 to 0.7) will occur with considerable frequency under the null hypothesis or under alternatives. On the other hand, with large sample sizes, a middle *p*-value provides evidence in favor of a null hypothesis. Accordingly, sample size plays a critical role in the interpretation of the results of significance tests. With adequate sample sizes, the results of significance testing should lead to the search for positive explanatory hypotheses, not the mere failure to reject a null hypothesis.

Criticism of Significance Tests from Other Fields

The significance testing issue has also been carefully reviewed in a volume edited by Morrison & Henkel (26, 27) entitled, *The Significance Test Controversy*. Several of the articles in this text are generally critical of use of statistical inference methods of any kind, particularly for observational data. The volume does additionally provide a lucid description of the testing debate as it has developed in the fields of sociology and psychology. In this text Selvin (42) and Rozeboom (40) provide the strongest critiques of significance testing, and both call for elimination of its use in their respective fields of sociology and psychology. Selvin expresses the somewhat restricted view that significance tests (for that matter, all statistical inference) should only be

Table 1 Hypothetical results: determination of sex

	Experience 1			Experience 2		
		Judgment of sex			Judgment of sex	
Category	Total	Correct	Incorrect	Total	Correct	Incorrect
Expected by chance	10	5	5	1000	500	500
Physician's judgment	10	6	4	1000	505	495
p		0.38			0.38	

utilized when an element of randomness has been introduced into the study. This randomness can be introduced through random sampling from a population, or by random assignment of individuals to specified treatment groups. These conditions are virtually never met in the field of sociology; therefore, strict application of this principle would all but eliminate the use of statistical inference techniques in this field. Skipper et al (43) critique the arbitrary nature of a level of significance, and particularly encourage reporting of precise p-values rather than the degenerate form of "significant versus nonsignificant." Labovitz (22) comments on the choice of significance level but additionally stresses the importance of other factors in statistical significance testing. He emphasizes the role of practical alternative hypotheses and the plausibility of specific alternatives. In addition, he reminds us of the role of power and sample size in statistical testing. He also distinguishes the objectives of developing hypotheses versus testing hypotheses. Gold (19) critiques Selvin's paper (42) and echoes an often heard theme, namely that p-values alone are insufficient summary statistics for data analysis and that estimators and standard errors of estimates must be reported. Kish (21) recognizes the problems with the uncritical application of statistical significance testing in sociology, but also notes that the criteria required by Selvin (42) would virtually never be met in sociological applications. This would leave the field without formal criteria for data analysis, a point strongly echoed by Davis (6). It is noteworthy that Fleiss (9) makes a similar point in discussing biomedical data analysis. Finally, Winch & Campbell (50) describe common mistakes in the reporting of significance tests. They discourage the use of language such as "proof" when describing results of hypothesis testing.

Chandler (5) describes the issues of confidence interval estimation and hypothesis testing as it relates to problems in psychology. As noted above, Rozeboom (40) is strongly critical of statistical hypothesis testing procedures, and encourages the elimination of the null hypothesis significance test (and statistical inference techniques) from most psychological research endeavors.

Issues in Statistical Significance Testing

The preceding authors have described a number of issues associated with the use of significance tests and with statistical inference generally for behavioral research. Since many behavioral studies are of the observational, nonexperimental type found in public health and epidemiologic research, a number of the points made in that literature warrant consideration here. There are really two general issues: statistical issues and philosophy of science issues.

With regard to statistical issues, there is first a question of sampling, namely how are the generated data relevant to the applicability of the test? If indeed an element of randomness occurs either through treatment assignment or random sampling, then there is relatively strong agreement that the use of confidence interval estimation and statistical significance tests is justified. Even, when randomness is present, however, some would criticize the use of statistical significance testing, particularly if results are reported in the degraded form of accept/reject the null hypothesis. We would agree that exact *p*-values should be reported and dichotomization avoided where possible.

The second issue regarding the use of significance tests relates to settings in which the sample is really an entire population. In this case some critics would argue that statistical significance tests are meaningless, for one is analyzing the entire population rather than a sample from the population. Others assert that the data could be viewed as having arisen from a larger universe, and the use of tests justified. Furthermore, in the absence of randomness, many investigators would argue that measurement errors still tend to be random, or that one could view the sample as having arisen from a hypothetical infinite population. Finally, one could construct an abstract hypothetical model from which assumptions such as random assignment of individuals to certain groups could be assumed to hold. Under this assumption of random assignment one could then generate the distribution for certain statistics that would describe the association between variables. This leads to well-defined rules for deciding whether associations are meaningful, although the use of the term "meaningful" must be carefully understood to apply to the context of this hypothetical model. In particular if the null hypothesis is rejected, the explanation may not be that the treatment is effective but that our assumptions are incorrect (e.g. selection bias, or errors of measurement as in the Fisher-Berkson example).

The third question relates to the meaning of statistical tests, particularly with regard to substantive versus statistical significance. It is important to note the points made by Berkson with regard to the role of sample size and how one may interpret meaningful differences or the absence of meaningful differences carefully in light of the sample size and power of the study.

As noted by several authors, the choice of significance levels is a rather arbitrary matter, and 0.05 and 0.01 seem to have taken a rather prominent role in data analysis. Ideally, the choice of alpha level should be related to the cost involved in the type of decision under consideration. These costs must also be balanced by the sample size and the power of the investigation. Of course, the arbitrary selection of significance level (0.05) or confidence level (95%) is an issue that affects confidence interval estimation and hypothesis testing equally.

Data editing and presentation can also affect the observed *p*-values or values of statistics used to determine the *p*-values. Some have cautioned that it is difficult to interpret results of publication articles in which the author may have collapsed data or may have chosen cut points for data summaries that may more favorably describe the results in light of the theories supported by the authors.

It is also important to recognize that statistical significance in no way determines causality. Associations that are important in a statistical sense must be assessed in light of other factors, especially scientific knowledge, and consistency of the results with data of a similar type. Although results must be interpreted in light of scientific theories, it is our view that having well-defined explicit statistical criteria for determining whether associations are meaningful and worthy of further explanation is crucial. This does not mean that a p-value of 0.05 is necessary. Depending upon the subject matter, size of the study, and state of knowledge in the field, a p-value of 0.10 or 0.15 could be deemed worthy of further investigation. This process involves the more difficult matter of interpreting results and incorporating them within existing theory. The results of a statistical significance test should be regarded as a first, and virtually never a last, step in the data analysis.

The philosophy of science issues associated with the use of significance testing relate to the logical issues associated with the use of tests and in particular the same issues raised by Berkson. Berkson questions the logic of rejecting a hypothesis for the occurrence of an event that can happen under the stated hypothesis. He does not recommend the abolition of significance testing in medical research, he merely recommends the careful study of p-values and the possible alternatives that are supported by the observed data. This to us seems to be sensible data analysis, and we would support this view very strongly.

Confidence Intervals and Statistical Significance Testing

If one is able to describe a particular parameter of interest, say theta, then it is generally quite meaningful to attempt to estimate the numerical value of theta and present a summary statistic reflecting the uncertainty of this estimate. There are, of course, a variety of ways in which this can be achieved. One such approach is based on confidence interval estimation in which an interval of values is presented and this interval reflects those values of the parameter that are in reasonable accord with the observed data. Reasonable accord can be translated into a specific number such as 95% or 99% confidence. Thus, confidence interval estimation (like testing) results in an arbitrary decision regarding confidence level. One way suggested around this approach is simply to present confidence intervals for a variety of numerical values, such as 95%, 99%, and 90%.

Significance tests can also be used to address questions associated with this parameter theta. In many contexts, for example in the assessment of differences between two treatment groups, it is important to answer the question regarding a specific value for this parameter. Statistical hypothesis testing is a direct way to address this question. As has been cited by a number of individuals, simply to report results of a hypothesis test in the form of accept or reject may not be informative. Given current computing capabilities,

p-value computations are easy to perform, and indeed appropriately rounded p-values should be reported.

In addition, the so-called p-value functions (35) to describe the relationship between the p-value and meaningful alternative hypotheses may be useful in some settings. To date, the utility of these p-value functions for actual epidemiologic data analysis has been demonstrated only for the simplest types of problems involving a single scalar parameter and a rather simple likelihood function. For multi-parameter problems, the complex nature of competing alterntive hypotheses will complicate the careful and systematic study of such p-value functions. Nevertheless, whenever possible and whenever reasonable alternative hypotheses can be set forth, such functions may be helpful in evaluating hypotheses.

In our view the controversy regarding the use of confidence interval estimation, p-values, and significance testing has been unrealistically slanted toward discussion of very simple situations, for example the analysis of a 2×2 table. In most situations numerous variables need to be considered, and some criteria are required to reduce the number of variables to a manageable number. The use of significance testing in such settings is reasonable, although a keen eye should be kept open for the somewhat arbitrary decisions associated with the choice of significance level for entering and deleting variables from the developing models. Significance testing offers a workable alternative in these settings when data reduction is required. This does not mean that an investigator is required to have a slavish adherence to such tests; it simply means that these tests do provide a reproducible and replicable set of decision criteria to which the data can be subjected. In the final section of this paper we describe an example that shows how significance tests can be used to simplify the data structure and reduce the length of the confidence intervals for the parameters that are most relevant to the question under study.

Example

The example is taken from the NHANES I Epidemiologic Followup Study (NHEFS) (24). The goal of NHEFS is to examine the relationship of baseline clinical, nutritional, and behavioral factors assessed in the first National Health and Nutrition Examination Survey (NHANES I:1971–1975) to subsequent morbidity and mortality. Data collection for the initial phase of followup took place between 1982 and 1984 and included tracing of all NHANES I participants, determining their vital status, conducting in-depth interviews with surviving participants or with proxies for those who were deceased or incapacitated, conducting selected physical measurements, obtaining facility records for stays in hospitals or nursing homes that occurred during the period of followup, and obtaining death certificates for decedents.

For this example we assume that the purpose of the analysis is to assess the effects of smoking and high blood pressure (systolic blood pressure of 160 mmHg or more) as measured at baseline on mortality after eight years of followup. The study population used here includes 4317 white men and women aged 55–74 at the NHANES I baseline examination, 955 of whom died during the followup period. Table 2 shows the basic data. Because the proportions *(p)* cover a wide range (from 3 to 55%), we use the logit scale, log $[p/(1\text{-}p]$, and the corresponding odds ratios, to express differences.

Suppose that one of the objectives is to estimate the joint impact (in the logit scale) of smoking and high blood presure on mortality relative to those with no risk factor. This effect can be estimated for each of the eight age-sex combinations shown in Table 2. Thus, we could present eight confidence intervals (CIs) as in Figure 1 and consider the job done. It is apparent, however, that there is substantial overlap among these CIs and that they are rather wide, with an average length of 1.8 (sixfold relative odds).

We can shorten the CIs by pooling data in several ways. For example, we can take a weighted average of the age-specific estimates, weighting by the inverse of the variances. This will reduce the eight CIs to two (males and females) with an average length of 0.82 (2.3-fold relative odds). This ap-

Table 2 Eight-year death rates by agc, sex, smoking, and systolic blood pressure (SBP): white persons aged 55–74 in 1971–1975

			Male			Female		
Age	Smoker	High SBP[a]	Dead	Alive	Percentage dead	Dead	Alive	Percentage dead
55–59 years	No	No	16	156	9.3	6	191	3.0
	No	Yes	4	32	11.1	3	52	5.5
	Yes	No	22	108	16.9	14	69	16.9
	Yes	Yes	8	20	28.6	4	25	13.8
60–64 years	No	No	25	137	15.4	18	181	9.0
	No	Yes	6	23	20.7	5	64	7.2
	Yes	No	29	62	31.9	5	70	6.7
	Yes	Yes	7	16	30.4	4	14	22.2
65–69 years	No	No	102	216	24.4	60	417	12.6
	No	Yes	37	86	30.1	54	203	21.0
	Yes	No	73	121	37.6	14	83	14.4
	Yes	Yes	32	43	42.7	11	30	26.8
70–74 years	No	No	93	187	33.2	70	272	20.5
	No	Yes	60	80	42.9	65	175	27.1
	Yes	No	56	52	51.9	15	38	28.3
	Yes	Yes	29	24	54.7	8	15	34.8

[a] Systolic blood pressure 160 mmHg or higher.

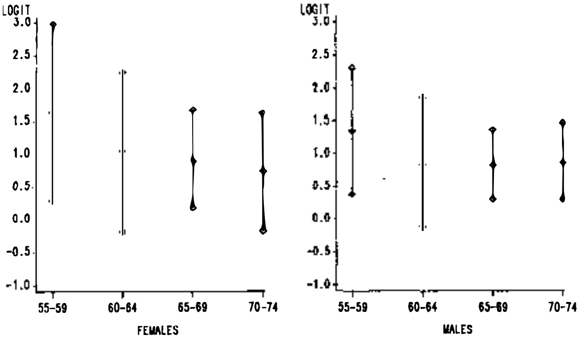

Figure 1 Odds ratios for 2 vs 0 risk factors with 95% confidence intervals.

proach, however, has two disadvantages. First, by comparing only hypertensive smokers to persons with neither risk factor present, the data from subjects with only one risk factor present are ignored. Second, pooling the data to reduce the eight age-sex specific estimates into two age-adjusted estimates does not provide a measure of the variation among the relative risks being pooled.

A more reasonable approach to the analysis of these data is to fit statistical models, assess the degree to which the models fit the data (using significance tests), and provide confidence intervals around the effect estimates of primary interest. The idea behind such model building is that the observed proportions in each cell are not necessarily good estimates of the probability of death, primarily because many of these proportions are based on small numbers. Model building provides "smoothed" estimates of the probability of death for every cell of the table. If the model is correct these estimates will be unbiased and have much smaller variances (and shorter CIs) than the observed proportions. Even if the model is incorrect and the estimates are biased, the variances of the smoothed estimates may be so much smaller than the variance of the observed proportions that the mean square errors [i.e. variance + $(\text{bias})^2$] of the smoothed estimates may be lower than those of the observed proportions.

In addition, the model building gives us insight into the structure of the data, allowing us to summarize the 32 death rates in Table 2 with fewer parameters that can be interpreted as the effects of the risk factors on mortality. The assessment of whether the model is correct is most easily carried out by using significance tests to determine whether the model provides an adequate fit to the data. In this example we illustrate these procedures by using weighted least squares estimates based on the GENCAT program (23). Other approaches are, of course, possible [e.g. maximum likelihood estimates based on the BMD-P logistic regression package (8) gave nearly identical results in this example].

The most general approach is to begin with a fully saturated model and eliminate interaction terms in a stepwise hierarchical manner. Table 3 shows the results. If we allow the deletion of any term with a *p*-value above 0.1, all interaction terms can be deleted. The resulting main effects model has a chi-square goodness-of-fit value of 21.88 with 25 degrees of freedom ($p = .643$). The estimates and confidence intervals based on this model are shown in Table 4. These results provide a succinct summary of the information contained in the much more detailed table with fewer and shorter confidence intervals than were available from the observed data. Furthermore, the *p*-values for the goodness-of-fit test and the stepwise deletion of terms provide an objective assessement of the degree to which the estimates in Table 4 adequately reflect the data.

The alternative to significance tests in this example is rather cumbersome. It is possible to generate CIs corresponding to the significance tests in Table 3. The CIs do have one advantage over the significance tests: They illustrate the magnitude of the interaction effects. Although this can be useful when examining two-way interactions, the number and complexity of the CIs become unwieldy in the case of three- and four-way interactions. For example, consider the model with the full age-sex-smoking interaction and a main effect for high blood pressure (goodness-of-fit $X^2 = 8.90$ with 15 degrees of freedom (df), $p = .883$). The odds ratios (OR) (relative to 55–59-year-olds) and CIs for the age-sex-smoking interaction with this model (corresponding to the X^2 test in line 8 of Table 3) are given below:

60–64:	4.49 (1.07, 18.84)
65–79:	3.47 (1.06, 11.33)
70–74:	3.08 (0.91, 10.41)

Table 3 Stepwise deletion of interaction terms from saturated model

Term deleted	df	X^2	*p*
Age-sex-smoking-SBP	3	4.11	.250
Age-smoking-SBP	3	0.81	.848
Age-sex-SBP	3	2.09	.553
Age-SBP	3	0.78	.855
Sex-smoking-SBP	1	0.26	.610
Smoking-SBP	1	0.12	.732
Sex-SBP	1	0.86	.353
Age-sex-smoking	3	5.43	.143
Age-scx	3	1.99	.575
Sex-smoking	1	1.59	.207
Age-smoking	3	4.71	.194

Table 4 Estimates for main effects model[a]

	Odds ratio	95% CI
Age (relative to 55–59)		
60–64 years	1.59	1.14, 2.20
65–69 years	2.65	2.01, 3.48
70–74 years	4.37	3.30, 5.77
Males versus females	2.09	1.79, 2.44
Smoker versus nonsmoker	1.86	1.57, 2.21
High versus normal SBP	1.44	1.23, 1.70
Sm•ker and high SBP versus neither	2.69	2.12, 3.40

[a] Goodness-of-fit chi square = 21.88 with 25 df, $p=.643$.

These parameters indicate that the male-female ratio of the smoking odds ratios for each age group 60–74 is 3–4.5 times the corresponding ratio for 55–59-year-olds. In order to better understand these estimates, it is helpful to calculate odds ratios and CIs for the smoking effect in each age-sex group based upon this model. These are shown in Figure 2. For each age group 60–74 the smoking OR for males is higher than for females but for 55–59-year-olds the reverse is true. The data therefore suggest that, although the smoking effect is relatively constant across ages for males, females 55–59 years of age have much larger relative odds for smoking than older females.

This supposition is confirmed from the significance tests by examining components of the 3 degrees of freedom X^2 test for the age-sex-smoking interaction. The X^2 for the age-sex-smoking interaction among those 60 and over is 0.35 with 2 df so that the x^2 of 5.43 for the age-sex-smoking interaction for all ages is almost entirely due to differences in the sex-smoking intcraction betwcen 55–59 vcrsus 60–74 (X^2=5.08, 1 df, p=.024).

These results suggest that a model with three different smoking effects (males, females 55–59 and females 60–74) will provide an adequate fit. Fitting this model to the data gives a goodness-of-fit X^2 of 11.97 with 23 df (p=0.971). The pertinent parameter estimates for this model are shown in Table 5. The inclusion of this age-sex-smoking interaction adds complexity to the interpretation of results. Under this model there are three different smoking effects: 2.00 for males, 3.64 for females 55–59, and 1.34 for females 60–74. There are also three sex ratios: 2.01 for nonsmokers, 1.10 for smokers 55–59, and 3.00 for smokers 60–74. Finally, there are two sets of age effects. One set, for male and female nonsmokers, shows a steady increase in deaths with increasing age (odds ratios of 1.9, 3.2, and 5.3 relative to 55–59-year-olds). The other set, for female smokers, shows odds ratios 0.7, 1.2, and 1.9 for each age group relative to 55–59-year-olds. The decline in the death rate from age 55–59 to age 60–64 seems rather implausible (note that the CI for

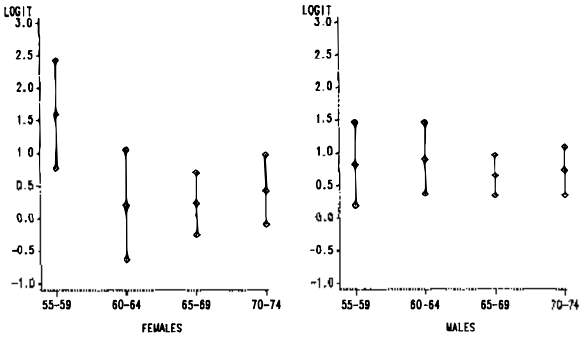

Figure 2 Odds ratios for smokers vs nonsmokers by sex and age with 95% confidence intervals.

the odds ratio for female smokers 60–64 relative to 55–59 includes values up to 1.3). It is, of course, impossible to know which of these models (Table 4 or 5) is closer to the true situation. Note, however, that all three confidence intervals in Table 5 for the smoking odds ratio overlap the corresponding CI in Table 4. In general, the isolation of single-degree-of-freedom components from an overall chi-square suffers from the dangers of multiple comparison problems: The chances of making a Type I error increase substantially. By setting a high *p*-value for deletion of interactions we run the risk of emphasizing differences in effects that merely reflect outliers in the particular data set being analyzed. By setting a low *p*-value for deletion, on the other hand, we risk missing potentially important differences in effects among subpopulations [and, in most studies, the power to detect interactions is low (20)]. The choice depends on the purpose of the analysis and the consequences of each type of error.

In summary, this example shows how both CIs and significance tests complement each other. Significance tests are valuable to determine which interaction terms can be deleted from a model and to assess the goodness-of-fit of the model. The reduced model in turn can be used to generate fewer and shorter CIs. CIs based on reduced models are preferable to significance tests because they provide information about the range of magnitude of the effects of interest. Using only one of these methods to analyze the data set would not provide the same degree of insight into the data structure that results when both are used.

Summary

The question of whether statistical significance testing should be used for the analysis of public health and epidemiologic data has received considerable attention in recent years. In this paper we have described some of the arguments for and against the use of hypothesis testing for the analysis of

Table 5 Estimates from model with modified age-sex-smoking interaction[a]

	Odds ratio	95% CI
Age (relative to 55–59)		
Male and female nonsmokers		
60–64 years	1.92	1.35, 2.74
65–69 years	3.18	2.35, 4.31
70–74 years	5.25	3.86, 7.15
Female smokers		
60–64 years	0.71	0.38, 1.31
65–69 years	1.17	0.65, 2.12
70–74 years	1.93	1.07, 3.51
Males versus females		
Nonsmokers	2.01	1.67, 2.41
Smokers 55–59 years	1.10	0.62, 1.98
Smokers 60–74 years	3.00	2.14, 4.22
Smoker versus nonsmoker		
Males	2.00	1.63, 2.45
Females 55–59 years	3.64	2.01, 6.58
Females 60–74 years	1.34	0.97, 1.85
High versus normal SBP	1.45	1.23, 1.70
Smoker and high SBP versus neither		
Males	2.89	2.23, 3.75
Females 55–59 years	5.27	2.84, 9.78
Females 60–74 years	1.94	1.34, 2.79

[a] Goodness-of-fit chi square = 11.97 with 23 df, $p = .971$.

biomedical data. In addition, we have reviewed the literature from related fields, in particular sociology and psychology, in which similar discussions have taken place within the last 30 years. Many of the significance testing criticisms in these scientific fields have been raised in the more recent discussions taking place in the biomedical field.

We present an example that emphasizes the use of both confidence interval estimation and significance testing. The example is particularly pertinent because it represents a more complex problem than has generally been discussed by critics of significance testing. Much of the discussion on this topic has focused on simple data analysis, such as the analysis of a 2x2 table or problems involving simple linear regression. Most epidemiologic data are far more complicated and warrant the use of both confidence interval estimation and significance testing for statistical analysis.

Both of these techniques have no doubt been misused in the analysis of data. These misuses may have arisen from a lack of understanding of the role of statistical methods in data analysis and the choice of such methods for data analysis.

If used prudently and judiciously, significance testing can help reduce the number of variables involved in a statistical analysis, thereby resulting in shorter confidence intervals for the models presented. Both significance testing and confidence interval estimation can serve and have served very useful functions for the analysis of public health and biomedical data.

Literature Cited

1. Bailar, J. C., Mosteller, F. 1988. Guidelines for statistical reporting in articles for medical journals. *Ann. Int. Med.* 108:266–73
2. Bandt, C. L., Boen, J. R. 1972. A prevalent misconception about sample size statistical significance, and clinical importance. *J. Periodont.* 43(3):181–83
3. Berkson, J. 1942. Tests of significance considered as evidence. *J. Am. Stat. Assoc.* 37:325–35
4. Browner, W. S., Newman, T. D. 1987. Are all significant P values created equal? *J. Am. Med. Assoc.* 257(18): 2459–63
5. Chandler, R. E. 1970. The statistical concepts of confidence and significance. See Ref. 26, pp. 213–16
6. Davis, J. A. 1970. Some pitfalls of data analysis without a formal criterion. See Ref. 26, pp. 91–94
7. Diamond, G. A., Forrester, J. S. 1983. Clinical trials in statistical verdicts: Probable grounds for appeal. *Ann. Int. Med.* 98:385–94
8. Dixon, W. J., Brown, M. B., Engelman, L., et al 1985. BMDP statistical software 1981. Los Angeles: Univ. Calif. Press
9. Fleiss, J. L. 1986a. Significance tests have a role in epidemiologic research: Reactions to A. M. Walker. *Am. J. Public Health* 76(5):559–60
10. Fleiss, J. L. 1986b. Confidence intervals versus significance tests: Quanititative intrepretation. *Am. J. Public Health* 76(5):587
11. Fleiss, J. L. 1986C. Response to A. M. Walker. *Am. J. Public Health* 76(8): 1033–34
12. Fisher, R. A. 1925. *Statistical Methods for Research Workers* London: Oliver & Boyd. (Subsequent eds. 1928, 1930, 1932, 1934, 1936, 1938, 1941, 1944, 1946, 1950, 1954, 1958).
13. Fisher, R. A. 1935a. *The Design of Experiments*. London: Oliver & Boyd. (Subsequent eds. 1937, 1942, 1947, 1949, 1951)
14. Fisher, R. A. 1935b. The logic of inductive inference. *J. R. Stat. Soc.* 98 (Part 1):39–54
15. Fisher, R. A. 1936. Uncertain inference. *Proc. Am. Acad. Arts Sci.* 71(4):245–58
16. Fisher, R. A. 1955. Statistical methods in scientific induction. *J. R. Stat. Soc. Ser. B* 17:69–78
17. Freiman, J. A., Chalmers, T. C., Smith, H., et al. 1978. The importance of data, the type II error, and sample size in the design and interpretation of the randomized control trial: A survey of 71 "negative" trials. *N. Engl. J. Med.* 299: 690–94
18. Gardner, M. J., Altman, D. G. 1986. Confidence intervals rather than *p*-values—Estimation rather than hypothesis testing. *Br. Med. J.* 292:746–50
19. Gold, D. 1970. A critique of tests of significance. See Ref. 26, pp. 107–8
20. Greenland, S. 1983. Tests for interaction in epidemiologic studies: A review and a study of power. *Stat. Med.* 2:243–51
21. Kish, K. 1970. Some statistical problems in research design. See Ref. 26, pp. 127–41
22. Labovitz, S. 1970. Criteria for selecting a significance level: A note on the sacredness of .05. See Ref. 26, pp. 166–71
23. Landis, R. J., Stanish, W. M., Koch, G. S. 1976. A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). *Comput. Programs Biomed.* 6:196–231
24. Madans, J. H., Kleinman, J. C., Cox, C. S., Barbano, H. E., Feldman, J. J., Cohen, B., Finucane, F. F., Cornoni-Huntley, J. 1986. 10 years after NHANES I: Report of initial followup, 1982–84. *Public Health Rep.* 101(5): 465–73
25. Mills, J. L., Reed, G. F., Nugent, R. P., et al. 1985. Are there adverse effects of periconceptional spermicide use? *Fertil. Steril.* 43:442–46
26. Morrison, D., Henkel, R., eds. 1970. *The Significance Test Controversy, Chicago: Aldine*
27. Morrison, D. E., Henkel, R. E. 1970. Significance tests in behavioral research: Pessimistic conclusions and beyond. See Ref. 26, pp. 305–12

28. Moses, L. E., Emerson, J. D., Hosseini, H. 1984a. Analyzing data from ordered categories. *N. Engl. J. Med.* 311:442–48
29. Moses, L. E., Emerson, J. D., Hosseini, H. 1984b. Reply to Poole, et al. *N. Engl. J. Med.* 311(21):1383
30. Neyman, J., Pearson, E. S. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Ser. A* 231:289–337
31. Neyman, J., Pearson, E. S. 1932. The testing of statistical hypotheses in relation to probabilities a priori. *Proc. Cambridge Philos. Soc.* 29:492–516
32. Pearson, E. S. 1955. Statistical concepts in their relation to reality. *J. R. Stat. Soc. Ser. B* 17:204–7
33. Pearson, K. 1911. Probability that two independent distributions of frequency are really samples from the same population. *Biometrika* 8:250–54
34. Poole, C., Lanes, S., Rothman, K. J. 1984. Analyzing data from ordered categories. *N. Engl. J. Med.* 311(21): 1382
35. Poole, C. 1987a. Beyond the confidence interval. *Am. J. Public Health* 77(2): 195–99
36. Poole, C. 1987b. Mr. Poole's response. *Am. J. Public Health* 77(10):1356–57
37. Rennie, D. 1978. Vivé la difference ($p < 0.05$). *N. Engl. J. Med.* 299: 828
38. Rothman, K. J. 1978. A show of confidence. *N. Engl. J. Med.* 299(24):1362–63
39. Rothman, K. J., Yankauer, A. 1986. Editor's note (Letters). *Am. J. Public Health* 76:587–88
40. Rozeboom, W. W. 1970. The fallacy of the null hypothesis significance test. See Ref. 26, pp. 216–30
41. Salsburg, D. S. 1985. The religion of statistics as practiced in medical journals. *Am. Statistician* 39(3):220–23
42. Selvin, H. C. 1970. A critique of tests of significance in survey research. See Ref. 26, pp. 94–106
43. Skipper, J. K., Jr., Guenther, A. L., Mass, G. 1970. The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. See Ref. 26, pp. 155–60
44. Thompson, W. D. 1987a. Statistical criteria in the interpretation of epidemiologic data. *Am. J. Public Health* 77(2):191–94
45. Thompson, W. D. 1987b. On the comparison of effects. *Am. J. Public Health* 7794):491–93
46. Vaisrub, N. 1985. Manuscript review from a statistician perspective. *J. Am. Med. Assoc.* 253(21):3145–47
47. Walk, A. 1950. *Statistical Decision Functions*. New York: Wiley
48. Walker, A. M. 1986a. Reporting the results of epidemiologic studies. *Am. J. Public Health* 76(5):556–58
49. Walker, A. M. 1986b. Significance tests represent consensus and standard practice. *Am. J. Public Health* 76(8):1033
50. Winch, R. F., Campbell, D. T. 1970. Proof? No. Evidence? Yes. The significance of tests of significance. See Ref. 26, pp. 199–208